A Clustering Technique for Value-Range Queries with Area Size Constraints

Ruixin Yang, Kwang-Su Yang, Menas Kafatos
Center for Earth Observing and Space Research (CEOSR)
School of Computational Sciences
George Mason University
Fairfax, VA, 22030, U.S.A.
{ryang, kyang, mkafatos}@gmu.edu

Abstract— **This paper describes a method for value-range queries under an area size condition. The query conditions are: the average value is in a certain range, the number of missing values is less than a specific percentage, and the contiguous area size is larger than a given limit. The method involves multi-level hierarchical clustering with depth-first strategy, convex hull and point-in-polygon algorithms. Examples of this method with Earth science data are given. The paper also discusses the future plan, especially, the development of a web-based prototype of the above algorithms and the implementation with a database management system.**[1]

I. Introduction

Data from the Earth Observing System (EOS) and other Earth observing platforms have been increasing explosively these days. An efficient method is necessary to search massive data sets. Content-based browsing or searching[1] is one of the promising methods by which a user searches data based on data values before subsetting and/or accessing data sets. The content-based queries provide users with areas in which the parameter values are in a given range. For scientists, the areas within the range of average value for a given parameter may be attractive. In addition, scientists may be concerned with areas larger than a certain size which preferably do not contain a large proportion of missing data points. These three conditions constitute the constraints in this study.

Yang et al.[2] came up with a fast method for searching high resolution data points within a certain value range for a given parameter. The method deals with queries with value range, focusing on the reduction of searching time based on a two level pyramid data model[1]. Clusters of low resolution cells are constructed, using the similarity among the histograms of the low resolution cells. There are other work on value-range queries, area constraints, and clustering in the cases with existential obstacles (e.g., [3], [4], [5], [6], [7]). The current work is dealing with both averaging value constraint and an existential constraint, the area size.

II. Method

Clustering techniques[8], [9] have been widely studied and used in many disciplines. The specific clustering method we present here is based on the hierarchical depth-first clustering technique. First, a large number of spatial clusters at the first level are generated, considering the data points within a value range and possibly removing spatial outliers. Three constraints are checked in the order of size constraint, missing data proportion constraint, and average value constraint. Assuming that the area of a cluster at a certain level is checked. If it does not meet the first constraint, the area is discarded. Then, the next cluster at the same level, branching from the same cluster at the previous level, or the next cluster at the previous level is checked. If it is satisfying the first constraint, the area is then clustered to examine whether its smaller areas within it may meet all constraints. If it meets all three constraints, it is chosen for display. The generation of clusters and constraint checking are based on depth-first strategy. Details of our method are reported in [10].

The constraint checking is carried out by using convex hull and point-in-polygon algorithm[11], [12]. Since the constraints are related to area, the areas need to be constructed in some way. Convex hull makes the area compact because it generates the smallest convex domain containing data points[12]. The areas generated by convex hull are more reasonable than those by KD-tree[13]. The latter may contain unnecessary areas which result from rectangular ranges used in KD-tree. Point-in-polygon algorithm is computationally intensive. To reduce its computation time, only data points within x- and y-range of convex hull are taken into account. Convexity property of convex hull makes it easy to determine whether a data point is located inside the convex hull. Data points inside the convex hull may or may not lie within the value range. Data points and missing data points inside the convex hull are counted and the average value is calculated for constraint checking.

The method is to find reliable, feasible areas satisfying the given constraints which are the size limit of an area with less than a specified proportion of missing data points and the average value within a specified range. The method is

[1] This work has been partially supported by NASA ESTO (NAG5-8607 and NAG5-10177).

based on the depth-first clustering strategy branching from infeasible areas to feasible areas. Since there is a large number of areas satisfying the size constraint, a feasible method should be developed based on computational time in real situation. The number of grid data points is used to represent the size of an area. For example, the size of an area is constrained such that it is not smaller than a certain number of grid data points. Since these points are distributed both uniformly and regularly, the number of the grid data points is approximately equivalent to the size of the area over which they are located.

Earth science data have a characteristic such that as the data points are closer, their attribute values tend to be more similar. Based on this spatial characteristic, the data points within a value range also appear to be closer together. The hierarchical agglomerative clustering technique applies to those data points. While the areas generated by the clustering technique contain the data points within the value range and those outside the value range, they consist of many data points within the value range. If the areas consist of two groups of data points, for which the values are far away from the value range, and are satisfying the constraints, they are not the areas that one wants to find. That is why the areas generated by our method are called reliable.

Figures 1 illustrates how the algorithm is working with four levels. The number of clusters at the first level is large and one of the clusters is chosen here for the illustration. The number of clusters at the second, the third and the fourth level is 3 , 2, and 2, respectively. There are four types of data points shown in Figure 1. The symbol $\bullet$ denotes a data point within the value range, the symbol $\triangle$ a data point above the maximum of the value range, the symbol ∇ a data point below the minimum of the value range, and the symbol $\times$ a data point with missing value. The hierarchical agglomerative clustering technique is applied to data points within the value range. The convex hull is drawn with a thicker line showing one cluster at the first level. The first number in the sequence identifying a cluster starting from the left denotes the cluster at the first level, the second number from the left denotes the cluster at the second level, the third number from the left denotes the cluster at the third level, and the fourth number from the left denotes the cluster at the fourth level. For example, Cluster 1.2.1.2 means the second cluster at level 4, belonging to the first cluster at level 3, belonging to the second cluster at level 2, belonging to the first cluster in level 1.

Assuming that the size constraint for data points is 7, that the missing proportion constraint is 0.2, and that the average constraint of an area is specified. Figure 1 shows clusters up to four levels. The area of the first cluster at the first level, Cluster 1.0.0.0, shown here is not supposed to be satisfying the average constraint. Three clusters at the second level are generated using the data points within the value range, $\bullet$, and they are Cluster 1.1.0.0, Cluster 1.2.0.0, and Cluster 1.3.0.0. Since the data points, $\bullet$, are considered, the areas of the three clusters at second level exclude most of data points outside the value range but inside the convex hull of Cluster 1.0.0.0. The candidate three areas appear to be reliable because the areas consist of many data points within the given value range. The area of Cluster 1.1.0.0 is satisfying the first two constraints and is supposed to be satisfying the average constraint. Next, the convex hull of Cluster 1.2.0.0 is constructed. The area is satisfying the first two constraints. However, if it turns out not to meet the average constraint, the clustering technique is applied to data points within the value range inside this convex hull. Two clusters, which are Cluster 1.2.1.0 and Cluster 1.2.2.0, are then generated. The area of Cluster 1.2.1.0 is satisfying the first two constraints. If it does not meet the average constraint, the clustering technique is applied to the data points inside this convex hull and clusters at the next level are then generated. They are Cluster 1.2.1.1 and Cluster 1.2.1.2. The area of Cluster 1.2.1.1 is satisfying all constraints because the number of data points is 8 and the values of all data points in this convex hull are within the value range. The area of Cluster 1.2.1.2 does not meet the size constraint so that it is discarded. Next, Cluster 1.2.2.0 is considered. The area of Cluster 1.2.2.0 is satisfying the first two constraints, since the number of data points is ten and there is no missing data point. If the area is satisfying the average constraint, the area is chosen. Finally, Cluster 1.3.0.0 is considered. Since the area of this cluster does not meet the size constraint, the area is discarded. All clusters generated have been considered so that the algorithm stops at this point.

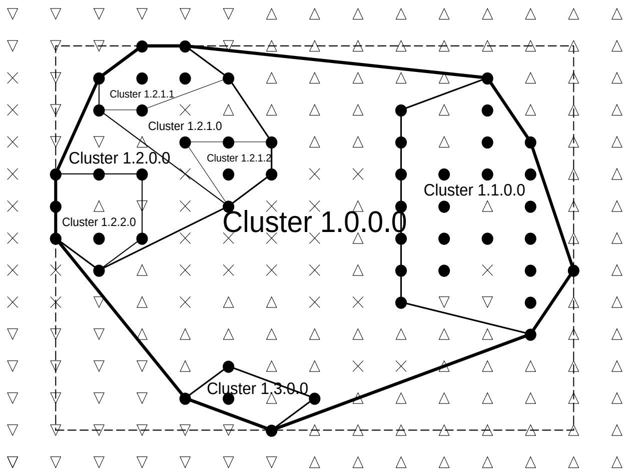

Fig. 1. Illustration of how the algorithm is working

III. Examples

We use a statistical analysis and graphics software, Splus, for implementation. The number of clusters at each

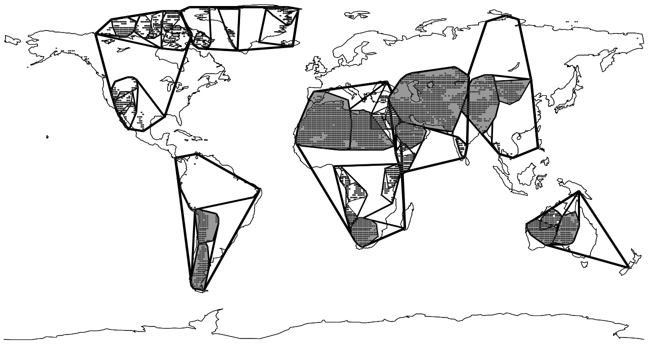

Fig. 2. Areas for NDVI data with average [0.04,0.26] and size limit with 100 data points and less than 15 percent of missing data points

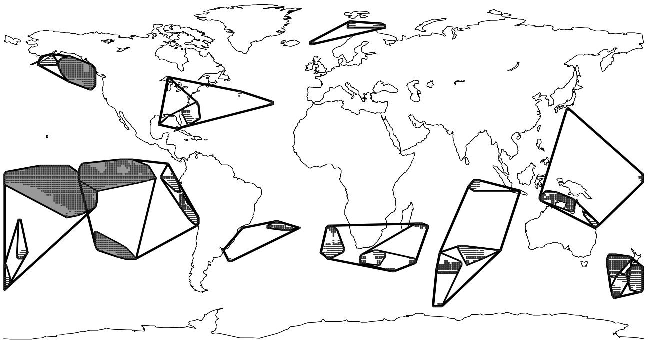

Fig. 3. Areas for SSTA data with average $[1.0, 4.0]^oC$ and size limit with 100 data points and less than 5 percent of missing data points

level, the minimum number of data points, the threshold percentage of missing points, and the value range are control parameters.

We use three sample data sets from CIDC (Climatology Interdisciplinary Data Collection) CD-ROM set[14]. The data are NDVI (Normalized Difference Vegetation Index), SSTA (Sea Surface Temperature Anomaly) and Surface Skin Temperature covering global land, global ocean, and globe respectively. The data sets are monthly mean values with 1 × 1 degree spatial resolution. We randomly pick the time, August 1981 for NDVI, January 1987 for both SSTA and Surface Skin Temperature.

Figure 2 is the example of NDVI. The size constraint is 100 data points, which provides areas containing not smaller than 100 data points. The average value range is taken to be between 0.04 and 0.26. Since the NDVI range between 0.04 and 0.26 is small, it is expected that the areas to be found will be low vegetation areas such as deserts. Since NDVI data cover global land, it is preferable that the areas do not include oceans. From this point of view, the minimum missing proportion is set to 0.15. That is, any area to be found should not contain the missing data points more than 15 percent of the total data points in that area. When the missing proportion is small, the areas have more chance of not meeting that constraint. In this case, these areas need to be further clustered. The number of clusters at the first level is 10 and the numbers at the rest of levels are all 2. Since the data points within the value range are over-plotted after finding the areas, the isolated data points are shown inside/outside convex hulls. Figure 2 shows the areas with gray color which contain 100 data points, have average NDVI values between 0.04 and 0.26, and have not more than 15 percent of missing data points. There are only seven clusters shown at the first level because each of other three clusters at the first level has a few number of data points so that they are eliminated for purposes of display. Eleven areas satisfying all constraints are found, which are convex hulls filled with gray levels. Areas found at each level are filled with dark gray color. Clustering proceeds up to 6 levels to find these areas. One area is found at the second level, four areas at the third level, and six areas at the fourth level. As expected, the low vegetation or desert areas cover the Middle East, some parts of China, northern Africa with the Sahara Desert and southern part of Africa, Australia, and Chile.

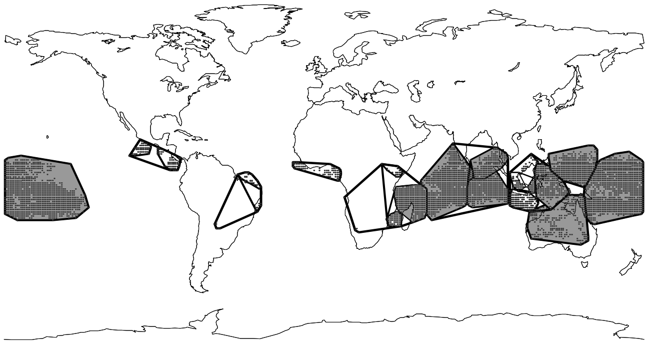

Fig. 4. Areas for Surface Skin temperature data with average $[301, 303]^oK$ and size limit with 100 data points and less than 5 percent of missing data points

Figure 3 shows the areas with gray color which contain more than 100 data points, have average SSTA values between 1.0^oC and 4.0^oC, and have not more than 5 percent of missing data points. There are many clusters with points on which the value range condition is satisfied. However, there are only four areas satisfying all conditions. Two larger areas are in tropical Pacific. The third is in northern part of the Pacific near the North America coast. The last area is also in Pacific but in the southern part.

Figure 4 shows the areas with gray color which contain more than 100 data points, have average Surface Skin Temperature values between 301^oK and 303^oK, and have not more than 5 percent of missing data points. Nine areas satisfying all constraints are found. Almost half of them are found at the first level, and others are found on either the second level or the third level. The areas satisfying the given conditions cover tropical areas from Indian Ocean until the eastern Pacific Ocean including most of Australia.

IV. Future Works

The technique described in this work shows promising results. We plan to extend the work in three directions. The first is to use MODIS data to test the algorithm since the volume of MODIS data is much larger. The second and a major task is to build a web-based prototype to make such constrained value range queries available to web users. To do this, we plan to follow some standards such as GML (Geography Markup Language) and SVG (Scalable Vector Graphics). That is, we put a output layer on the system to make the output (various areas) in GML. Then, we transfer the GML into SVG for displaying on web browsers. We plan the GML-SVG path instead of direct SVG path because we expect third party software will be available for displaying information in GML format.

To have a workable prototype, we may need a database management system to support the indexing system. Therefore, our third potential action is to study the usage and efficiency of a database management system such as Oracle for implementing the prototype. In particular, Oracle Spatial Option will be focused to study the potential for such a prototype.

References

[1] Z. Li, X. S. Wang, M. Kafatos, and R. Yang, "A Pyramid Data Model for Supporting Content-based Browsing and Knowledge Discovery," in *Proceedings of the 10th International Conference on Scientific and Statistical Database Management* (M. Rafanelli and M. Jarke, eds.), pp. 170–179, IEEE, Computer Society, 1998.

[2] R. Yang, K. Yang, M. Kafatos, and X. Wang, "Value Range Queries on Earth Science Data via Histogram Clustering," in *Interim Proceedings of International Workshop in Temporal, Spatial and Spatio-Temporal Data Mining, TSDM2000, Lyon, France, Lecture Notes in Artificial Intelligence, 2001* (K. Hornsby and J. F. Roddick, eds.), Springer, 2001.

[3] C. H. Cheng, "A Branch-and-Bound Clustering Algorithm," *IEEE Transactions on Systems, Mans, and Cybernetics*, vol. 25, pp. 895–898, 1995.

[4] M. N. Murty and G. Krishna, "A Computationally Efficient Technique for Data Clustering," *Pattern Recognition*, vol. 12, pp. 153–158, 1980.

[5] A. K. H. Tung, J. Hou, and J. Han, "COE: Clustering with Obstacles Entities, A Preliminary Study (PDF)," in *Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000.*

[6] A. K. H. Tung, J. Hou, and J. Han, "Spatial Clustering in the Presence of Obstacles," in *Proc. 2001 Int. Conf. on Data Engineering (ICDE'01), Heidelberg, Germany, April 2001.*

[7] A. K. H. Tung, , J. Han, L. V. S. Lakshmanan, and R. T. Ng, "Constraint-Based Clustering in Large Databases," in *Proc. 2001 Int. Conf. on Database Theory (ICDT'01), London, U.K., Jan. 2001.*

[8] B. S. Everitt, *Cluster Analysis.* John Wiley & Sons, 1993.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, pp. 264–323, 1999.

[10] K.-S. Yang, R. Yang, and M. Kafatos, "A Feasible Method to Find Areas with Constraints Using Hierarchical Depth-First Clustering," in *Proceedings of the 13th International Conference on Scientific and Statistical Database Management* (L. Kerschberg and M. Kafatos, eds.), pp. 257–262, IEEE, Computer Society, 2001.

[11] J. O'Rourke, *Computational Geometry in C.* Cambridge University Press, 1994.

[12] F. P. Preparata and M. I. Shamos, *Computational Geometry.* Springer-Verlag, 1985.

[13] H. Samet, *Applications of Spatial Data Structures.* Addison-Wesley Publishing Company, 1990.

[14] H. L. Kyle, J. M. McManus, S. Ahmad, and et al., *Climatology Interdisciplinary Data Collection, Volumes 1-4, Monthly Means for Climate Studies.* NASA Goddard DAAC Science Series, Earth Science Enterprise, National Aeronautics & Space Administration, NP-1998(06)-029-GSFC, 1998.